Poster: Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks

Fatemehsadat Mireshghallah[1], Kartik Goyal[2], Archit Uniyal[3]
Taylor Berg-Kirkpatrick[1], Reza Shokri[4]
[1] University of California San Diego, [2] Toyota Technological Institute at Chicago (TTIC)
[3] Panjab University, [4] National University of Singapore
[fatemeh, tberg]@ucsd.edu,
kartikgo@ttic.edu,archituniyal.dev@gmail.com,reza@comp.nus.edu.sg

***Abstract*—The wide adoption and application of Masked language models (MLMs) on sensitive data (from legal to medical) necessitates a thorough quantitative investigation into their privacy vulnerabilities – to what extent do MLMs leak information about their training data? Prior attempts at measuring leakage of MLMs via membership inference attacks have been inconclusive, implying potential robustness of MLMs to privacy attacks. In this work, we posit that prior attempts were inconclusive because they based their attack solely on the MLM's model score. We devise a stronger membership inference attack based on likelihood ratio hypothesis testing that involves an additional reference MLM to more accurately quantify the privacy risks of memorization in MLMs. We show that masked language models are extremely susceptible to likelihood ratio membership inference attacks: Our empirical results, on models trained on medical notes, show that our attack improves the AUC of prior membership inference attacks from 0.66 to an alarmingly high 0.90 level, with a significant improvement in the low-error region: at 1% false positive rate, our attack is $51\times$ more powerful than prior work.**

***Index Terms*—Fairness, Privacy, Natural Language Processing, Language Modeling**

I. Introduction and Approach

BERT-based models with Masked Language Modeling (MLM) Objectives have become models of choice for use as pre-trained models for various Natural Language Processing (NLP) classification tasks and have been applied to diverse domains such as disease diagnosis, insurance analysis on financial data, sentiment analysis for improved user experience, etc. Given the sensitivity of the data used to train these models, it is crucial to conceive a framework to systematically evaluate the leakage of training data from these models, and limit the leakage. The conventional way to measure the leakage of training data from machine learning models is by performing membership inference attacks, in which the attacker tries to determine whether a given sample was part of the training data of the target model or not. These attacks expose the extent of memorization by the model at the level of individual samples. Prior attempts at performing membership inference and reconstruction attacks on masked language models have either been inconclusive [1], or have (wrongly) concluded that memorization of sensitive data in MLMs is very limited and these models are more private than their generative counterparts (e.g., autoregressive language models) [2], [3]. We hypothesize that prior MLM attacks have been inconclusive because they rely solely on overfitting signals from the model under attack – i.e, the target model. More specifically, they use the target model's loss on each individual sample as a proxy for how well the model has memorized that sample. If the loss is lower than a threshold, the sample is predicted to be a member of the training set. However, the target model's loss includes confounding factors of variation – for example, the intrinsic complexity of the sample – and thus provides a limited discriminative signal for membership prediction. This scheme has either a high false negative rate (with a conservative threshold) – classifying many hard-to-fit samples from the training set as non-members, or a high false positive rate (with a generous threshold) – failing to identify easy-to-fit samples that are not in the training set.

Reference-based likelihood ratio attacks, on the other hand, when applied to certain probabilistic graphical models and classifiers, have been shown to alleviate this problem and more accurately distinguish members from non-members. In such attacks, instead of the loss of the model under attack, we look at the ratio of the likelihood of the sample under the target model and a reference model trained on samples from the underlying population distribution that generates the training data for the target model. This ratio recalibrates the test statistic to explain away spurious variation in model's loss for different samples due to the intrinsic complexity of the samples. Unlike most other models (e.g., generative models), however, computing the likelihood of MLMs is not straightforward. In this paper, we propose a principled framework for measuring information leakage of MLMs through likelihood ratio-based membership inference attacks and perform an extensive analysis of memorization in such models, as shown in Figure 1. To compute the likelihood ratio of the samples under the target and the reference MLMs, we view the MLMs as energy-based probabilistic models over the sequences. This enables us to perform powerful inference attacks on conventionally non-probabilistic models like masked language models.

We empirically show that *our attack improves the AUC from* 0.66 *to* 0.90 on the ClinicalBERT-Base model, and achieves *a true positive rate (recall) of* 79.2% (for false positive rate of 10%), which is a substantial improvement over the baseline with 15.6% recall. This shows that, contrary to prior results,

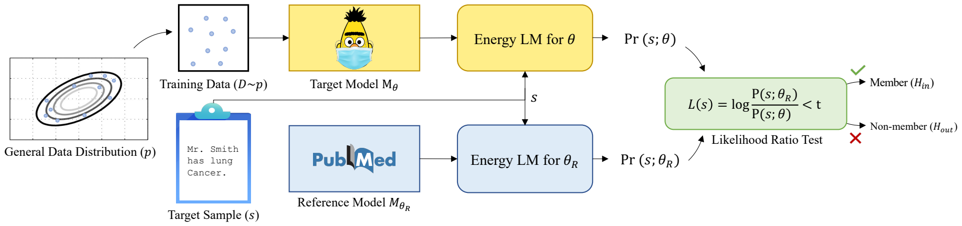

Fig. 1: Overview of our attack: to determine whether a target sample s is a member of the training data ($D \sim p$) of the target model (M_θ), we feed it to the energy function formulation of M_θ so that we can compute $\Pr(s;M_\theta)$, the probability of s under M_θ. We do the same with a reference model M_{θ_R} which is trained on a disjoint data set from the same distribution as the training data. Then, we compute likelihood ratio $L(s)$, and based on this ratio and a given test threshold t, we decide if s is a member of D (H_{in}) or not (H_{out}).

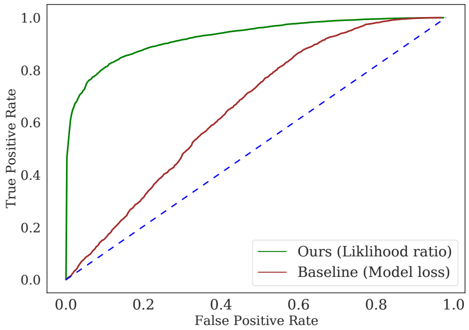

Fig. 2: The ROC curve of sample-level attack on Clinical-BERT with MIMIC used as non-member. Green line shows our attack and the red line shows the baseline loss-based attack. The blue dashed line shows AUC=0.5 (random guess). This figure corresponds to the results presented in the first column of Table I.

TABLE I: Overview of our attack on the ClinicalBERT-Base model, using PubMed-BERT as the reference. Sample-level attack attempts to determine membership of a single sample, whereas patient-level determines membership of a patient based on all their notes. The MIMIC and i2b2 columns determine which dataset was used as non-members in the target sample pool.

		Sample-level		Patient-level	
	Non-members	MIMIC	i2b2	MIMIC	i2b2
AUC.	(A) Model loss	0.662	0.812	0.915	1.000
	(B) Ours	0.900	0.881	0.992	1.000
Prec.	(A) w/ μ thresh.	61.5	77.6	87.5	100.0
	(A) w/ Pop. thresh.	61.2	79.6	87.5	92.5
	(B) w/ Pop. thresh.	88.9	87.5	93.4	92.5
Rec.	(A) w/ μ thresh	55.7	55.8	49.5	49.5
	(A) w/ Pop. thresh.	15.6	39.0	49.5	100.0
	(B) w/ Pop. thresh.	79.2	69.9	100.0	100.0

masked language models are significantly susceptible to attacks exploiting the leakage of their training data. The code and files needed for reproducing our results and building on our work are publicly available as part of the ML Privacy Meter tool[1].

II. Evaluation

We evaluate our proposed attack on a suite of masked clinical language models, following [1]. We compare our attack with the baseline from the prior work that relies solely on the loss of the target model [3], as this is the only way privacy in MLMs is evaluated. To further highlight the extent of the privacy risk, we show that in low error regions, where the inference attack has very small false positive rates, the attack has a large true positive rate – *at* 1% *false positive rate, our attack is* 51× *more powerful than the prior work*, as shown in Figure 2. Table I shows the metrics for our attack and the baseline's on both sample and patient level, with held-out MIMIC-III and i2b2 medical notes used as non-member samples. Here, the target model under attack is ClinicalBERT-base. Figure 2 shows the ROC curve of our attack and the baseline, for the sample-level attack with MIMIC-III held-out data as non-members samples. The table shows that our method significantly outperforms the target model loss-based baseline [3], which threshold the loss of the targets the model based on either the mean of the training samples' loss (μ), or the population samples' loss. Our attack's improvement over the baselines is more apparent in the case where both the members and non-members are from MIMIC-III, which is actually the harder case where the baselines perform poorly, since in and out samples are much more similar and harder to distinguish if we only look at the loss of the target model. Our attack, however, is successful due to the use of a reference, which helps magnify the gap in the behavior of the target model towards members and non-members, and is, therefore, better at teasing apart similar samples.

References

[1] Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron Wallace, "Does BERT pretrained on clinical notes reveal sensitive data?," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, June 2021, pp. 946–959, Association for Computational Linguistics.

[2] Thomas Vakili and Hercules Dalianis, "Are clinical bert models privacy preserving? the difficulty of extracting patient-condition associations," in *Proceedings of the AAAI 2021 Fall Symposium on Human Partnership with Medical AI : Design, Operationalization, and Ethics (AAAI-HUMAN 2021)*, 2021, number 3068 in CEUR Workshop Proceedings.

[3] Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu, "Membership inference attack susceptibility of clinical language models," 2021.

[1] https://github.com/privacytrustlab/ml_privacy_meter/tree/master/ml_privacy_meter/attack/mlm_mia